View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

对 DEA 聚类分析方法的一种改进*

李 果　王应明
(厦门大学自动化系　361005)

摘　要　本文提出一种新的 DEA—最优分割聚类分析方法。该方法以 DEA 相对效率评价为基础，通过最优分割法寻找最优分割点，从而达到聚类的目的。聚类结果准确、可靠，且能得到比较清晰、满意的结果，因而是一种既科学又有效的聚类分析方法。文章最后以一个实例说明该方法的具体应用。

关键词　聚类分析　DEA 方法　最优分割法

1　引言

文献[1]从相对效率角度提出了一种 DEA 聚类分析方法，该方法以 DEA 有效为聚类标准，通过逐步寻找 DEA 有效单元的方法从而达到聚类的目的。该方法特别适宜处理同时具有多个输入和多个输出指标的样本品的聚类分析，而且该方法是以 DEA 有效为聚类标准，比较科学和客观。在文献[2]中用 DEA 方法对 1991年合肥地区的工业行业进行了效率评价，并给出了按不同效率水平的分类。计算结果表明，当决策单元(行业)过多时，类数也可能较多，分类结果过细(比如 1991年合肥地区的工业行业分成了 12类)，而如果对相近的类作主观综合，又易造成人为给定聚类标准(比如文献[2]中按第一轮评价相对效率为 0.78和 0.4进行分类综合)，使得分析过于片面和主观化，从而使聚类结果缺乏说服力。

针对上述不足，本文在运用 DEA 方法进行相对效率评价的基础上，采用最优分割的方法进行聚类，该方法因为是用最优分割法计算出最优分割点来聚类，因此比人为给定聚类标准更为科学和客观。

2　DEA—最优分割法聚类原理

最优分割法是对有序样品进行聚类的方法。该方法用来分类的依据是离差平方和。设样品依次是 $x_1, x_2, \cdots, x_n$ (每个是 m 维向量)，最优分割法的步骤为：

(1) 定义类的直径。设某一类 G_{ij} 是 $\{x_i, x_{i+1}, \cdots, x_j\}$，$j > i$，它们的均值记成 $\bar{x}_{ij}$，

$$\bar{x}_{ij} = \frac{1}{j-i+1}\sum_{l=i}^{j} x_l$$

G_{ij} 的直径用 $D(i,j)$ 表示，常用的直径是

$$D(i,j) = \sum_{l=i}^{j} (x_l - \bar{x}_{ij})^T (x_l - \bar{x}_{ij})$$

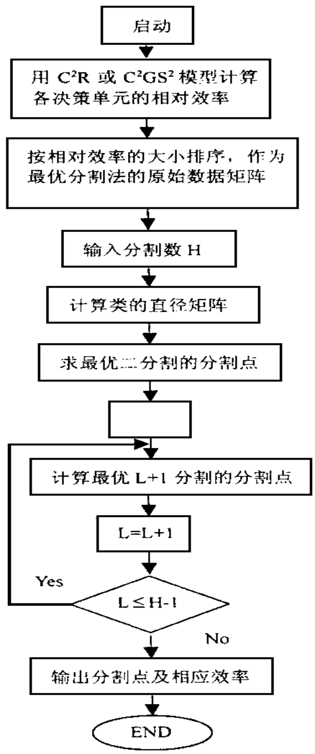

图 1

(2) 定义目标函数。将 n 个样品分成 k 类，设某一种分法是：

$$P(n,k): \{x_{i_1}, x_{i_1+1}, \cdots, x_{i_2-1}\}, \{x_{i_2}, x_{i_2+1}, \cdots, x_{i_3-1}\}, \cdots, \{x_{i_k}, x_{i_k+1}, \cdots, x_n\}$$

或简记成

$$P(n,k): \{i_1, i_1+1, \cdots, i_2-1\}, \{i_2, i_2+1, \cdots, i_3-1\}, \cdots, \{i_k, i_k+1, \cdots, n\}$$

* 收稿日期：1999-01-12
基金项目：国家自然科学青年基金资助项目 (79600020)

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.c

其中分点 $1 = i_1 < i_2 < \cdots < i_k < i_{k+1} = n$

定义这种分类的目标函数为:

$$e[p(n, k)] = \sum_{j=1}^{k} D(i_j, i_{j+1} - 1)$$

当 n, k 固定时, $e[p(n, k)]$越小表示各类的离差平方和越小, 分类是合理的。

(3) 精确最优解的求法

容易验证有如下的递推公式:

$$e[p(n, 2)] = \min_{2 \le j \le n} \{D(1, j-1) + D(j, n)\} \quad (1)$$

$$e[p(n, k)] = \min_{k \le j \le n} \{e[p(j-1, k-1)] + D(j, n)\} \quad (2)$$

当我们要分成 k 类时, 首先找 j_k 使(2)式达到极小, 即

$$e[p(n, k)] = e[p(j_k - 1, k-1)] + D(j_k, n)$$

于是 $G_k = \{j_k, j_k + 1, \cdots, n\}$, 然后找 j_{k-1} 使它满足

$$e[p(j_k - 1, k-1)] = e[p(j_{k-1} - 1, k-2)] + D(j_{k-1}, j_k - 1)$$

得到类 $G_{k-1} = \{j_{k-1}, \cdots, j_k - 1\}$, 类似的方法得到所有类 $G_1, \cdots, G_k$, 这就是我们欲求的最优解。

用 DEA— 最优分割法对决策单元进行效率评价, 并进行聚类分析, 只需先对各决策单元用 C^2R 模型或 C^2GS^2 模型计算出相对效率, 再将所有相对效率按大小排序, 然后给出聚类数目, 再运用最优分割法自动求得分割点。这样, 我们就可以得到比较满意的聚类结果。其过程可用图 1的框图表示。

3 应用举例

我们对文献[2]中 1991年合肥地区工业产业 30个不同行业的投入产出效率进行了实例分析。用 DEA 方法中的 C^2R 模型进行相对效率评价的结果见表 1的倒数第 2列, 假如我们希望将 30个不同的工业行业聚成 5类, 应用本文最优分割法聚成 5类的结果见表 1的最后 1列。

表 1 工业行业 DEA— 最优分割法聚类结果分析

工业行业	全部职工年平均人数(人)	全部资金(万元)	工业总产值(万元)	相对效率	分类结果
烟草加工业	2374	31297	37159.8	1.0000	1
工艺美术品制造业	2010	1365	2370	0.8810	1
医药工业	3526	10541	20774.8	1.0000	1
橡胶制品业	6799	16160	26712.8	0.8464	1
电子及通讯设备制造业	6856	30774	44467.6	0.8852	1
建材及其它非金属矿物采选业	3179	1414	1982.3	0.7113	2
缝纫业	5127	4188	6890.6	0.8348	2
造纸及纸制品业	3992	4323	6865.2	0.8129	2
金属制品业	10225	14161	15677	0.5854	2
电气机械及器材制造业	20218	69628	100547	0.7860	2
其它工业	1473	1209	1838.8	0.7717	2
食品制造业	14565	31912	31902.3	0.5138	3
纺织工业	35831	63228	48294.4	0.3972	3
家具制造业	4198	2684	2603.7	0.5014	3
印刷业	5266	10167	10152.9	0.5166	3
文教体育用品制造业	2750	2250	2531.8	0.5709	3
塑料制品业	5833	15821	15407.6	0.4959	3
建材及其它非金属矿物制品业	39198	27743	22044.8	0.4536	3
机械工业	50258	105563	93700.8	0.4571	3
交通运输设备制造业	25014	51693	44845.3	0.4447	3
自来水生产和供应业	1003	9588	2494.2	0.2014	4
饮料制造业	2916	11201	3619	0.1850	4
饲料工业	656	3150	1943.4	0.3884	4
皮革、毛皮及其制造业	2687	5360	3333.3	0.3155	4
木材加工及竹藤棕草制品业	2221	2738	1767.4	0.3299	4
炼焦、煤气及煤制品业	490	814	363.2	0.2264	4
化学工业	19959	62642	46080.5	0.3826	4
仪器仪表及其它计量器具制造业	3506	8476	6256.5	0.3777	4
黑色金属冶炼及压延加工业	20963	64346	21377.1	0.1709	5
有色金属冶炼及压延加工业	2348	9325	2266.3	0.1413	5

(下转 63页)

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

3 允许卖空最优组合证券模型

在允许卖空条件下, 最优组合的确定可采用如下非线性规划模型

模型(C)
$$\begin{cases} \max\Delta(X) = \max\limits_{X\neq 0} \dfrac{X^T AX}{X^T VX} \\ \text{s. t.}\ X^T e_n = 1 \end{cases}$$

求解模型(C)可利用文献[3]关于特征根的极值性质。

设 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_P$ 为对称阵 A 相对于正定阵 V 的特征根, 即 $\lambda_1, \lambda_2, \cdots, \lambda_P$ 是方程 $|A - \lambda V| = 0$ 的根, $t_1, t_2, \cdots, t_P$ 为对应于 $\lambda_1, \lambda_2, \cdots, \lambda_P$ 的 A 相对于 V 的单位特征向量, 则 $\max\Delta(X) = \max\limits_{X\neq 0} \dfrac{X^T AX}{X^T VX} = \lambda_1$, 且最大值在 $X = ct_1$ 时达到, c 为任意不等于零的实数。设 $t_1 = (t_{11}, t_{12}, \cdots, t_{1n})^T$, 对 $X = ct_1$ 利用约束条件 $X^T e_n = 1$, 即可得到模型(C)的最优解是 $X = (1/\sum\limits_{i=1}^{n} t_{1i})t_1$。

4 不允许卖空最优组合证券模型

这时最优组合的选择可采用如下非线性规划模型

模型(D)
$$\max\Delta(X) = \max\limits_{X\neq 0} \frac{X^T AX}{X^T VX}$$
$$\text{s. t.} \begin{cases} X^T e_n = 1 \\ X \geqslant 0 \end{cases}$$

模型(D)只是理论公式, 即使最优解存在也很难求出。实际应用中仍可按模型(B)获得一系列有效组合, 比较其投资效率, 以投资效率最大者为最优组合。

求解最优组合的步骤如下:

(1) 给出模型(B)的一组 $\_$。

据文献[4], 模型(B)有解的必要条件是: $\_{\min} \leqslant \_{0} \leqslant \_{\max}$, 其中 $\_{\min} = \min\{\_{1}, \_{2}, \cdots, \_{n}\}$, $\_{\max} = \max\{\_{1}, \_{2}, \cdots, \_{n}\}$, 我们将区间 $[\_{\min}, \_{\max}]$ 进行 m 等分(如 $m = 10$ 或 20), 各分点分别为 $\_^{(0)} = \_{\min}$, $\_^{(1)} = \_^{(0)} + L/m$, $\_^{(2)} = \_^{(0)} + 2L/m, \cdots, \_^{(i)} = \_^{(0)} + iL/m, \cdots, \_^{(m)} = \_{\max}$, 其中 $L = \_{\max} - \_{\min}$, 然后令模型(B)中 $\_{0} = \_^{(i)}$, $i = 0, 1, \cdots, m$。

(2) 对固定的 $\_{0} = \_^{(i)}$ 用改进的树形算法求模型(B)的最优解即有效组合 $X_B^{(i)}$。

先从 n 种证券组合出发(作为根结点), 然后按(1)式计算无非负约束模型(A)的最优解 $X_A^{(i)}$, 如 $X_A^{(i)}$ 无负分量就不再分支; 若有 k 个负分量, 则从原组合结构中剔除一种和负分量对应的证券, 这样做有 k 种可能, 这就形成了 k 个具有 $n-1$ 种证券组合, 它们作为根结点的 k 个分支(子结点), 构成了根结点后的第一层结点。将第一层中每个结点依照根结点的作法形成的所有分支结点构成了第二层, 再从第二层的每个结点展开下去, 直到不再具有可以分支的结点为止, 这就形成了一棵树形结构。该树的叶子结点处的无非负约束最优投资比例向量满足非负性, 我们再对每个叶子结点处的证券组合使用(2)式计算风险值, 值最小的叶子结点即为模型(B)的最优解 $X_B^{(i)}$。

(3) 获得模型(D)的近似最优解, 即近似最优组合。

若 $\max\limits_{0\leqslant i\leqslant m}\Delta(X_B^{(i)}) = \max\limits_{0\leqslant i\leqslant m} \dfrac{(X_B^{(i)})^T AX_B^{(i)}}{(X_B^{(i)})^T VX_B^{(i)}} = \Delta X_B^{(i_0)}$, 则 $X_B^{(i_0)}$ 就是我们要找的近似最优组合系数向量。

参 考 文 献

[1] 唐小我. 经济预测与决策新方法及其应用. 成都: 电子科技大学出版社, 1997.

[2] 马永开, 唐小我. 非负约束条件下组合证券投资决策方法的进一步研究. 预测, 1996(4).

[3] 方开泰. 实用多元统计分析. 上海: 华东师范大学出版社, 1989.

[4] 杨德权, 胡运权, 翟成强. 均值方差模型最优解中出现负分量的条件研究. 预测, 1997(5).

[5] 邹长贵, 欧阳植. 证券组合有效性研究及实证分析. 数量经济技术经济研究, 1996(5).

(上接 67 页)

通过以上的实际聚类分析可以看出, DEA—最优分割聚类分析法对 DEA 聚类法作了改进, 它适宜于处理多个输入多个输出指标单元的聚类分析, 是一种效果很好的聚类分析方法。

4 结束语

本文在 DEA 聚类方法的基础上, 提出了一种新的 DEA—最优分割聚类分析方法, 该方法提供了比仅用 DEA 聚类分析所能得到的更为清晰、满意的结果, 而且由于该方法不受任何人为主观因素的影响, 因而是一种科学而客观的聚类分析方法, 聚类结果准确、可靠, 不具有主观随意性。

参 考 文 献

[1] 王应明, 傅国伟. 运用 DEA 方法进行聚类分析. 控制与决策, 1993, 8(2): 86～90

[2] 华中生, 梁梁. 应用 DEA 方法对合肥地区工业行业效率的评价. 系统工程与电子技术, 1994, 16(11): 67～73

[3] 张尧庭, 方开泰. 多元统计分析引论. 科学出版社, 1982: 444～457

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net